

Evaluating Lexical Variant Generation to Improve Information Retrieval

Guy Divita¹, Allen C. Browne² and Thomas C. Rindflesch², Ph.D.

¹Management Systems Designers, Vienna VA, 22180

²National Library of Medicine, Bethesda, MD 20894

Techniques for managing lexical variation constitute an integral part of information retrieval systems. We report on a series of experiments aimed at evaluating LVG, a lexical variant management tool which addresses the particular problems involved in matching health related vocabularies to concepts in the Unified Medical Language System[®] (UMLS[®]) Metathesaurus[®]. Experiments conducted on data from the Large Scale Vocabulary Test indicate the effectiveness of this approach to managing biomedical information.

INTRODUCTION

Traditionally, information retrieval systems have used stemmers to manage lexical variation and thus enhance performance [1]. Several studies have investigated the effectiveness of various stemming techniques and have compared suffix removal to linguistically-based methods for abstracting away from inflectional variation [2-6]. This research indicates that, in general, stemming is useful, and that methods which produce actual words as stems are preferable to those which do not. This latter characteristic is particularly significant for the extension of these techniques to interactive query expansion or applications involving a multi-lingual environment, natural language processing systems.

This paper reports on experiments which investigate the components of a lexically-based variant generation system in the biomedical domain. The system described, LVG [7], is under development at the National Library of Medicine and is distributed as part of the resources accompanying the Unified Medical Language System (UMLS) Metathesaurus. The experiments were conducted on data from the Large Scale Vocabulary Test (LSVT) [8] and are part of ongoing research aimed at enhancing LVG's linguistically-based methods for accommodating the biomedical language encountered when mapping user input terms to relevant Metathesaurus concepts.

Managing Lexical Variation

All methods aimed at managing lexical variation abstract away from inflection in order to achieve aggressive pattern matching. In contrast to stemmers,

which achieve this goal by cutting off some number of characters from the ends of words [9] or by matching and replacing suffixes [2], LVG uses a large knowledge base (the SPECIALIST Lexicon [7]) to manage inflectional morphology. For example, in contrast to normal stemmers, when LVG is confronted with a word such as *thymus*, it does not remove the final *s* and produce the nonoccurring form *thymu*. This is prevented by the fact that *thymus* is a lexical entry and thus known to be the uninflected form of a noun.

LVG's approach to inflectional morphology is based on a set of rules [10] which both cooperate with and depend on the SPECIALIST Lexicon. These rules apply to English and account for singular and plural in nouns, tense in verbs, and comparative and superlative in adjectives and adverbs. To the extent that Greco-Latin inflectional variation is productive in modern English, it is also accommodated. Although the lexicon is large, with over 100,000 entries, it will never have complete coverage. For those words not in the lexicon, suffix morphology rules are applied. Exceptions to these rules are used as filters to prevent spurious forms from being generated.

The Structure of LVG

LVG has several architectural characteristics which were designed to accommodate a range of approaches to the management of English lexical variation in general and biomedical language in particular. We describe some of these characteristics before focusing on the experiments aimed at assessing their effectiveness.

Canonical Forms. A core LVG technique is to "uninflect" input terms to their base form. This process occasionally results in two legitimate uninflected forms for the same inflected input. For example, *left* uninflects to both *left* and *leave* reflecting its ambiguity as an adjective or verb. A technique to manage this ambiguity produces only one "canonical" base form for any given input term. The process of canonicalization precomputes all uninflected forms and then arranges these into classes composed of terms that could be expanded to the same inflected form. The

canonical form is an arbitrarily chosen member of this class and represents all the members of the class. For example, the terms *left*, *leave*, and *leaf* are all included in one such class, and the canonical form is *leaf*, the alphabetically first, and shortest member of the class.

LVG Flows. LVG has the capability to transform text in various ways which not only allows the management of English morphological phenomena but also can assist in accommodating other types of lexical variation encountered in the biomedical vocabulary. For example, simple transformations (or “flows”) can leave the input term untouched, convert all letters to lower case, remove punctuation, or sort words into ascending ASCII order. Additional flows uninvert input, transforming *Cancer, Lung* to *Lung Cancer*; remove stop words (such as *and*, *of*, and *the*); and remove genitive markers, changing *Down’s Syndrome* to *Down Syndrome*.

More interestingly, LVG can inflect terms, for example generating *sleeping*, *slept*, and *sleeps* from *sleep*, as well as uninflect, for example, *acute generalized tuberculoses* to *acute generalized tuberculosis*. (A variation on uninflection applies to words within terms and produces *acute generalize tuberculosis*.) LVG can also generate derivational variants (for example *medical* from *medicine*), enumerate known spelling variants, expand acronyms and abbreviations, conflate acronym and abbreviation expansions, and list known synonyms.

Each transformation can be used as the input to another transformation, and sequences of flows can be combined into a single complex flow. A special complex flow, the normalization flow, has been designed to address the most common variation encountered in accessing UMLS Metathesaurus terminology. This flow removes stop words, strips genitive markers, strips punctuation, lowercases, uninflects each term, maps each uninflected term to a canonical form, word order sorts the result. For example, the Normalization Flow transforms *Disorders of the Autonomic Nervous System* to *autonomic disorder nervous system*.

LVG Development. As part of ongoing research aimed at improving LVG, a number of problems associated with existing flows were identified. Dates and numbers were not handled effectively, and multi-word terms (with and without hyphens) posed particular problems. In addition, problems associated with canonicalization led us to seek alternative ways of dealing with ambiguous base forms.

Several flows, both simple and complex, were created to address these issues. One, a simple flow, removes punctuation not occurring in numbers. Several com-

plex flows were introduced to deal with problems associated with multi-word terms. These newly-devised flows involve issues concerning both multi-word terms and canonicalization

The normalization process often invokes look-ahead to deal with terms which might consist of a single word, hyphenated forms, or multiple words. Look-ahead is a technique frequently used in LVG to deal with such data; however, its efficacy is suspect. In order to evaluate look-ahead, several flows were added in order to deal with phenomena such as *breast-feeding* as a variant of *breastfeeding*.

We were also interested in pursuing an effective normalization flow which does not depend on canonical forms. An alternative would return both base forms in instances of ambiguity (for example, both *left* and *leave* as bases for *left*). In order to address this issue, two normalization complex flows were created not based on canonicalization, one with a previous approach to punctuation and another with an enhanced approach. A third flow addressing the phenomenon of ambiguous bases returns multiple bases if the input term is ambiguous and occurs in the lexicon, but returns a canonical base for input not in the lexicon. We were interested in determining whether any of these could match (or surpass) the effectiveness of the current normalization flow.

METHODS

We took advantage of data from the The Large Scale Vocabulary Test (LSVT) [8] in order to evaluate the effectiveness of the LVG components under development. The LSVT was conducted to determine whether existing health-related terminologies address vocabulary requirements in health care information systems and provides a valuable resource as a test collection for experiments evaluating lexical variation.

Sixty participants submitted over 40,000 terms to the test and through the use of a variety of tools determined whether their input term conceptually matched an existing UMLS Metathesaurus concept, discounting variation phenomena such as morphology, spelling, and synonymy.

The lexical variation experiments discussed in this paper were applied only to the LSVT data which represents a match between a participant’s input term and some concept from the Metathesaurus (21,472 terms). Two examples are the following, where an LSVT participant’s input term is followed by the matching Metathesaurus concept and its unique identifier (CUI). In addition to case differences, the first example involves

inversion, while the second has both inversion and inflectional variation.

GLOSSITIS,ATROPHIC

Atrophic glossitis

C0267044

nephrostomy tube

Tubes, Nephrostomy

C0184149

The particular value of this data for our experiments is that matches abstracting away from variation phenomena have been vetted by humans and can thus serve as a standard for evaluating lexical variation management techniques.

Experiments were conducted by submitting the LSVT data to several LVG flows, both simple and complex. We were particularly interested in assessing the effectiveness of complex normalization flows designed to enhance the treatment of punctuation, ambiguous base forms, and multi-word terms. For each flow tested, the output contained the LSVT input (term, matching Metathesaurus concept, and CUI) along with the output term as transformed by the flow. The 1997 Metathesaurus concept file [12] was sent through the same flow producing output consisting of transformed concepts and associated CUI's.

The transformed LSVT output terms and CUI's were then compared to the transformed Metathesaurus concepts and CUI's. If the CUI from the LVG output matched for both the LSVT terms and the Metathesaurus concepts, this indicated that LVG had made the same match as the LSVT participant in mapping input term to Metathesaurus concept.

The standard evaluation measures of recall and precision along with an f-score [10] which combines these metrics were used to score the effectiveness of each flow tested. On the basis of recall (R) and precision (P), the f-score is computed as:

$$2 (\beta + 1) PR / 2 (\beta P + R) = f$$

In this equation, the variable β weights the relative importance of R and P and was given a value of 1 in order to treat the two equally.

Testing to determine the significance of results has not yet been done. Parametric tests in information retrieval experiments are not valid because necessary underlying assumptions about the nature of the data are not met [12]. Nonparametric tests ([5 or 14], for example) need to be pursued; however, given the range of values in the scores for the various normalization techniques employed, the the general trends

seen here are likely to persist in additional experiments.

RESULTS

As a preliminary to examining the results seen for the innovative complex flows, we give (in Table 1) scores for the baseline (do nothing) and the crucial simple flows which support the complex normalization flows. In the table, a description of the flow is followed first by the LVG code for that flow, and then the evaluation metrics, recall, precision, and f-score; finally, relative processing time for the entire set of data is given as Unix User Time in hours, minutes and seconds. Note that the baseline, against which improvements are measured, has an f-score of .41

The results for the newly-devised complex normalization flows under inspection are given in Table 2, where the columns are as in Table 1. "RP" indicates the unenhanced flow for removing punctuation, while "RPE" reflects the enhanced version. "C" refers to canonicalization and "NC" to non-canonicalization. The rows in Table 2 compare the current normalization flow to the innovative normalization flows introduced above. The first row is the current flow, using canonicalization for ambiguous bases and unenhanced treatment of punctuation. The second row indicates results due to modifying the previous flow by adding look-ahead for multi-word terms and using enhanced punctuation removal. The next three rows investigate the specific interaction of alternative approaches to ambiguous bases and treatment of punctuation. "Norm (NC, PR)" uses the unenhanced punctuation removal, while "Norm (NC, PRE)" uses the enhanced version. Both reflect normalization without canonicalization (both bases are returned in cases of ambiguity). The next row in Table 2 ("Norm (NC & C)" reports the results of returning both bases for ambiguous input which occurs in the lexicon and returning a canonical form otherwise.

Although modest difference in recall and precision are seen in the various approaches to normalization (from .66 to .67 for recall and from .87 to .92 for precision), the f-scores are identical (.77) for the normalization flows tested, except for the flow which mixes the canonical and noncanonical treatment of ambiguous bases, which is slightly lower at .75.

As the last line in Table 2, we include the results from a normalization flow which includes the Porter stemmer. As reported elsewhere [5], the results from a linguistically-based approach such as LVG and those obtainable from a stemmer are comparable, with the former being slightly better.

Description	Flow	Rec.	Prec.	F-Score	Time
Baseline	fn	0.26	0.95	0.41	4:35.1
Lowercase	fl	0.51	0.94	0.66	4:53.0
Remove genitives	fg	0.26	0.95	0.41	4:33.5
Word order sort	fw	0.30	0.94	0.46	5:41.3
Uninflect words	fB	0.35	0.93	0.51	1:09:26.2
Remove stop words	ft	0.29	0.94	0.45	10:00.7
Remove punctuation	fp	0.27	0.94	0.42	4:40.0
Remove punctuation (enhanced)	fP	0.28	0.94	0.43	6:30.7

Table 1. Simple Flows

Description	Flow	Rec.	Prec.	F-Score	Time
Norm (C, RP) (Current)	ftgNt	0.67	0.91	0.77	1:38:13.2
Norm with lookahead (C,PRE)	flgPtBCw	0.67	0.89	0.77	1:38:56.6
Norm (NC, PR)	ftgplBw	0.66	0.92	0.77	1:20:46.6
Norm (NC, PRE)	ftgPlBw	0.66	0.92	0.77	1:20:02.2
Norm (NC & C)	flgPtBEw	0.66	0.87	0.75	1:55:52.0
Norm with Porter	ftgPlqw	0.67	0.87	0.76	13:28 .1

Table 2. Complex Normalization Flows

DISCUSSION

The results reported in Table 2 indicate that aggressive normalization of complex, multi-word biomedical terms greatly improves performance over the baseline (f-score increase from .41 to .77). However, the results for a number of alternative approaches to normalization indicate that the augmented techniques proposed did not dramatically enhance matching effectiveness. This investigation is nonetheless instructive.

The enhanced method for punctuation removal, while providing a slight improvement in isolation (Table 1), has no noticeable effect when combined into a complex normalization flow. Further, the results for including look-ahead in the normalization flow indicate that this computationally intensive technique for addressing multi-word terms is not cost effective, and it would not be advisable to add it to our general approach to normalization.

Another technique, canonicalization, which currently is part of the normalization flow, can be eliminated without degrading effectiveness. The noncanonical approach to ambiguous stems is preferable in that it is easier to maintain and is intuitively more satisfying than the canonical method.

The LVG methodology for management of lexical variation is knowledge intensive and based on linguistic principles. The results from the most aggressive combination of techniques available to LVG compare favorably to the those achievable by the Porter stemmer. However, in producing actual words as base forms (rather than nonoccurring stems), LVG can be readily incorporated into other applications, such as natural language processing systems aimed at providing interpretation of biomedical text ([15] for example).

References

1. Salton G, McGill MJ. *Introduction to Modern Information Retrieval*, McGraw-Hill, Inc, 1983.
2. Krovetz R. Viewing Morphology as an Inference Process. In Korfhage R, Rasmussen E and Willett P(ed.) *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 191-202, 1993.
3. Church KW. One Term or Two. In Fox E, Ingwerson P, and Fidel R (eds) *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 310-318, 1995.
4. Aronson AR. The Effect of Textual Variation on Concept Based Information Retrieval. In Cimino J(ed) *Proceedings of the AMIA Fall Symposium*, 373-377, 1996.
5. Hull, DA. Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1):70-84, 1996.
6. Tzoukermann E, Klavans JL, and Jacquemin C. Effective Use of Natural Language Processing Techniques for Automatic Conflation of Multi-Word Terms: The Role of Derivational Morphology, Part of Speech Tagging, and Shallow Parsing. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 148-155, 1997.
7. McCray AT, Srinivasan S, and Browne AC. Lexical methods for managing variation in biomedical terminologies. In Ozbolt JG (ed.) *Proceedings of the 18th Annual SCAMC*, 235-239, 1994.
8. McCray AT, Cheh ML, Bangalore AK, et al. Conducting the NLM/AHCPR Large Scale Vocabulary Test: A Distributed Internet Based Experiment. In Masys D(ed.) *Proceedings of the AMIA Fall Symposium*, 560-564, 1997
9. Porter MF. An Algorithm For Suffix Stripping. *Program*, 14(3):130-137, 1980.
10. SPECIALIST Morphology Documentation, now on-line at <https://umlsks.nlm.nih.gov/KSS/LVG/morph.html>.
11. Appelt DE, Israel D. Tutorial on Building Information Extraction Systems. *Fifth Conference on Applied Natural Language Processing, from the Association for Computational Linguistics*, 1997.
12. Spark Jones K. *Information Retrieval Experiment*, Butterworths & Co, Ltd, 1981.
13. UMLS Knowledge Sources, 9th Edition, January 1998, US. Department of Health and Human Services, National Institutes of Health, National Library of Medicine.
14. Wilbur WJ. Non-parametric significance tests of retrieval performance comparisons. *Journal of Information Science*, 20(4):270-284.
15. Aronson AR, Rindfleisch TC, and Browne AC. Exploiting a large thesaurus for information retrieval. *Proceedings of RIAO 94*, 1994:197-216.